DOCUMENT RESUME

ED 139 214 EC 100 789

AUTHOR Halpern, Andrew S.
TITLE Principles and Practices of Measurement in Career Education for the Handicapped.
PUB DATE Apr 77
NOTE 34p.; Paper presented at the Annual International Convention, The Council for Exceptional Children (55th, Atlanta, Georgia, April 11-15, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Career Education; *Evaluation Criteria; *Handicapped; Legislation; *Measurement Techniques; Norm Referenced Tests; *Performance Criteria; Performance Tests; *Testing

ABSTRACT

Discussed is testing in the field of career education for the handicapped, with emphasis on four major topics: applied performance testing, criterion validity studies, product vs. process measurement, and criterion vs. norm-referenced measurement. The author reviews some political considerations relevant to this area of testing. (IM)

COLLEGE OF EDUCATION
CENTER ON HUMAN DEVELOPMENT

REHABILITATION RESEARCH AND TRAINING CENTER
IN MENTAL RETARDATION

UNIVERSITY OF OREGON, EUGENE

PRINCIPLES AND PRACTICES OF MEASUREMENT IN CAREER EDUCATION FOR THE HANDICAPPED

Andrew S. Halpern
April 15, 1977

1

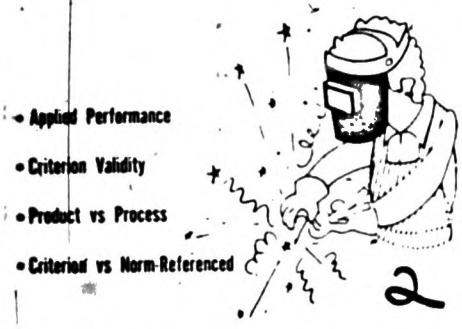
Introduction

1/ There are many measurement methods, both formal and informal, that have evolved within the field of career education for the handicapped. These methods include the interpretation of historical records and interviews, third-party evaluations, direct observations, and direct testing of behavior.

2/ This presentation will focus entirely upon testing, which is the most ubiquitous form of measurement in the field. Four major topics will be discussed:

- applied performance testing;
- criterion validity studies;
- product vs. process measurement;
- criterion vs. norm-referenced measurement.

3/ The presentation will conclude with some recent political considerations that are relevant to the field.



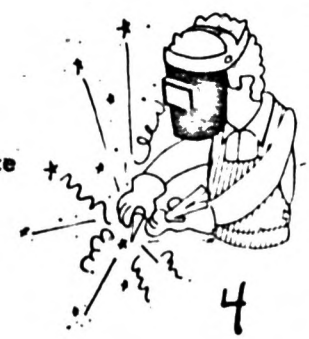
2 3

POLITICS

3

Applied Performance Testing

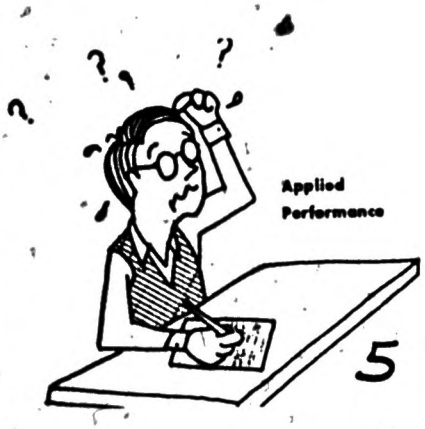
4/ In recent years, there has been a strong plea from several sources to focus our test development efforts upon "criterion sampling" (McClelland, 1973; Lavisky, 1975; Kulman, et al., 1975). If one is interested in predicting whether or not a person will succeed as a welder, why bother measuring the speed and accuracy with which he places pegs in holes?



Applied Performance

4

One of the more adament sources of this appeal can be found in the literature on applied performance testing (Lavisky, 1975). At its most basic definitional level, applied performance testing is accomplished when a desired behavior is measured <u>in the context</u> in which performance is expected. One step removed from contextual relevance would be measurement of the desired behavior in a <u>simulated</u> context. A second step removed would be measurement of <u>knowledge about</u> the desired behavior.



5/ In some instances, cognitive behavior or knowledge is an end in itself. When this occurs, measurement of knowledge can itself be construed as applied performance.

6/ Some of the work samples that have been developed for vocational evaluation of the handicapped are good examples of applied performance testing. The Vocational Evaluation and Work Adjustment Association offers the following definition of a work sample (Kulman, et al., 1975), as:

> a well defined work activity involving tasks, materials, and tools which are identical or similar to those in an actual job or cluster of jobs . . . . The work sample should simulate the complete range of work activities of which a particular job or occupational group is comprised. (p. 55)

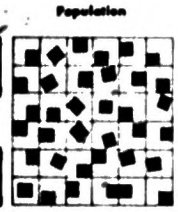
Not all work samples are as clearly related to criterion behaviors as this definition would suggest. Some involve criterion sampling only in a rather abstract sense, since the tasks are performed in a context far removed from an actual vocational setting. The appeal of applied performance testing lies not only in the common sense desirability of measuring what <u>really</u> is of interest, but also in the assumption that measurement of criterion behavior has much clearer relevance for instructional intervention than measurement of some surrogate for the criterion.

This philosophy of diagnostic/prescriptive measurement and intervention has almost become a religion in today's educational community, particularly among educators of the handicapped. There is much to be said for this position. Before totally joining the bandwagon, however, it is worth mentioning some of the precautions that were recently raised by a proponent of this position (Joselyn, 1975):

- Large amounts of time and energy are often required for the production and measurement of behavioral objectives.
- No single list of objectives is likely to serve the needs of many students and teachers.

- Easy-to-measure objectives may get chosen only because of this technical advantage, resulting in a neglect of more relevant but harder to measure objectives.
- If accountability is a major consideration, teachers may choose some objectives only because they are easily achievable.
- Focusing on precise educational objectives tends to detract from consideration of broad educational goals, which often leads to an assumption that existing goals are adequate.

10/ As one examines this list of precautions, one major factor emerges as the cause of many of the problems. Life is simply too complex to specify all the precise behavioral criteria that are desired outcomes of our interventions or the desired behavioral dimensions for our selection and classification decisions. Measurement, therefore, will almost always require a sampling of behavior, which raises the dual questions of the generalizability and the pertinence of ~~our findings~~ test scores to related content and to other contexts.

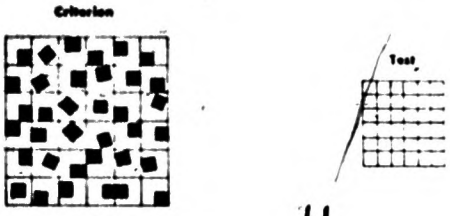


10

Criterion Validity Studies

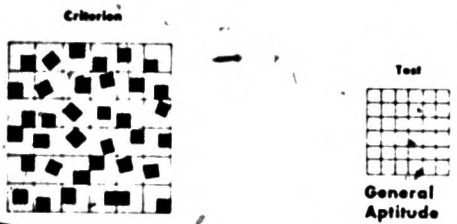
11/ Whenever we depart from criterion measurement, or measure only a sample of criterion behavior, we are

forced to raise the question of criterion validity. Most tests are subject to this requirement. Of the many that exist in the field of career education, I have chosen three that seem to be somewhat representative. One examines the relationship between general aptitudes and vocational criteria in a non-handicapped population. Another examines the relationship between work samples and vocational criteria in a handicapped population. The third examines the relationship between knowledge and pre-vocational criteria in a mentally retarded population.



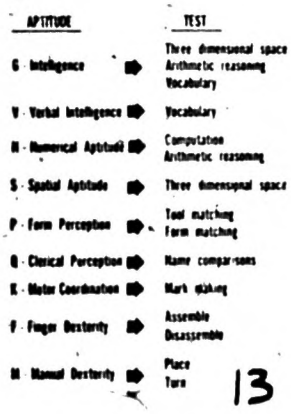
11

12/ General Aptitude Test Battery. The General Aptitude Test Battery, or GATB, is probably the most highly researched vocational evaluation instrument in existence. It is used primarily for the purpose of classification, to help place individuals in vocational training programs or jobs.



12

The battery consists of 12 tests yielding nine aptitude scores. The scores and their corresponding aptitudes are shown in Figure 1. /13 (PAUSE)



13

14/ The criterion validity of the GATB for different jobs or occupations is determined through the development of "validity norms." This process entails a number of well defined steps, beginning with the selection of an "experimental sample" consisting of people in a given

Validity Norms

14

occupation of preparing for a given occupation. The GATB is then administered to this sample, in order to identify those aptitudes with high means and low standard deviations. Criterion measures are next obtained, usually involving either production records or performance evaluations by supervisors. The final step in the analysis involves dichotomizing the relevant aptitudes, and then experimenting with multiple cutoffs within each aptitude in order to discover which arrangement maximizes the number of people who are either high scorers on both the aptitudes and the criterion or low scorers on both the aptitudes and the criterion. An example from the GATB Manual (Manual for the . . ., 1970) illustrates this process for the job of case worker.

Case Worker

G - Intelligence
V - Verbal Intelligence
N - Numerical Aptitude
Q - Clerical Perception

15

15/ The experimental sample for this study consisted of 106 case workers. After identifying G, V, N, and Q as potentially relevant aptitudes, the multiple cutoff analyses suggested that only G, V, and N be retained. The relationship between test scores and criterion performance is shown in Table 1. / 16

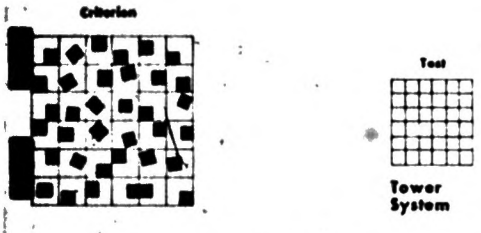
As this table shows, the obtained phi coefficient was indeed statistically significant. But what can we say of its practical significance? 17/ Twenty-one percent

Relationship Between Test Norms and Dichotomized Criterion - Case Worker N = 106

	Non-Qualifying Test Scores	Qualifying Test Scores	Total
Good Workers	20	54	74
Poor Workers	18	14	32
Total	38	68	106

$\phi = .28$ $P < .01$

16, 17, 18



19

of those who earned qualifying test scores turned out to be poor workers.[18] More potently, over 50% of those who did not earn qualifying test scores turned out to be good workers. How confident could we be, on the basis of this study, in using GATB scores to advise a person as to whether or not to become a case worker?

TOWER System.[19] Our second example of a criterion validity study comes from the literature on work sample evaluation with handicapped people. The TOWER System, developed at the Institute for the Crippled and Disabled in New York City, is the oldest and best known work sample evaluation system that has been developed for use with handicapped people (Testing, Orientation and Work Evaluation in Rehabilitation, 1974). It consists of 94 work samples that have been clustered into 14 occupational areas (ICD Rehabilitation and Research Center, 1974).

In 1967, a report was published describing a fairly comprehensive attempt to assess the criterion validity of the TOWER System (Rosenberg, 1967). In this study, relationships were examined between work sample evaluation scores, the "intermediate" criterion of supervisor OR ratings of performance during training, and the "ultimate" criterion of acquisition and retention of a job.

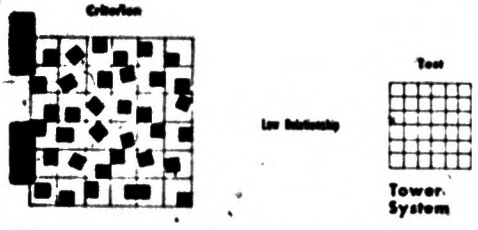
The primary predictor variables of interest were the "performance rating" and the "quality rating" that are derived from each work sample evaluation. Both ratings used a five-point scale, summarizing the evaluator's prediction on the likelihood of successful training and subsequent employment in occupations represented by the work sample.

When work sample evaluation was completed on the experimental sample, four possible types of disposition occurred:

- the client received vocational training in a trade area;
- the client received training in unskilled workshop activities;
- the client received direct placement at a job;
- the client was closed as not feasible.

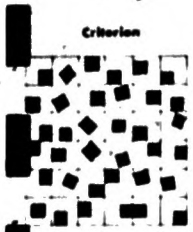
Subsequent analyses involved only members of the first three of these groups.

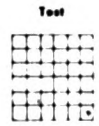
20/ The project was completely unsuccessful in demonstrating criterion validity for the TOWER System. For the most part, very low relationships were found between the predictor and the criterion variables. The author of the project report offered several explanations for these findings:



- Not being able to randomly assign subjects within each disposition greatly restricted the range of predictor scores within each disposition. The restricted range, in turn, set statistical limits on the possible magnitude of the correlation coefficients.
- Inter-rater reliabilities were not established for either the predictor or the criterion variables. Both sides of the equation were believed to be somewhat deficient.
- The administration procedures in the TOWER System were not fully standardized, requring many clinical judgments on potentially different behavioral samples.
- The quality of training provided to clients varied in an unknown manner.
- The follow-up procedures were inadequate and produced limited data.

21/ Social and Prevocational Information Battery.





Social & Prevocational Information Battery

21

The third example of a criterion validity study is derived from our own work at the Rehabilitation Research and Training Center in Mental Retardation at the University of Oregon. This series of studies examines the relationship between knowledge and criterion behavior in several social and prevocational domains.

During the past five years, our Center has been involved in the articulation of a measurement strategy and the creation of a standardized series of tests known as the Social and Prevocational Information Battery (SPIB). The SPIB was originally developed for use with mildly retarded people (Halpern, Raffeld, Irvin, & Link, 1975a; Irvin & Halpern, In press), and has been available in published form since September, 1975 (Halpern, Raffeld, Irvin, & Link, 1975b). A major revision of the SPIB for use with moderately retarded people has recently been completed and is available now (Irvin, Halpern, & Reynolds, 1977; Irvin, Halpern, & Reynolds, In press).

22/ The domains measured by the SPIB relate to five broad areas of adult community adjustment: employability, economic self-sufficiency, family living, personal habits, and communication. Within these five broad areas, nine separate tests have been developed, measuring the examinee's knowledge of job search skills, job related behavior, banking, budgeting, purchasing, home management, health care, hygiene and grooming, and functional signs (survival reading).

23/ Extensive field testing produced 277 items, either true/false or picture selection in format, which are

orally administered to examinees in order to neutralize the impact of differential reading ability.[24] Each of the nine tests contains approximately 30 items that can be administered separately, requiring 10 to 15 minutes for completion.

[25] Items were selected in accordance with a domain sampling model, whereby each domain or area to be tested was further specified hierarchically into 54 content areas and 180 sub-content areas across the nine SPIB domains. The sub-content areas then served as a blueprint for generating test items for possible inclusion in the battery.

Three studies have been conducted to examine the relationship between SPIB performance and applied performance.[30] In order to estimate predictive validity, a sample of vocational rehabilitation clients was rated by rehabilitation counselors on 29 behaviors in the broad areas covered by the SPIB. This same sample had been tested on the SPIB one year earlier, just prior to graduation from high school.

[31] The concurrent validity of the SPIB was also estimated with a second sample of vocational rehabilitation clients. These clients were tested on the SPIB

and rated by their counselors within a three-month period of time.

32 The concurrent validity of SPIB-T was estimated within a sample of residents from community facilities. The applied performance scale, in this study, consisted of 87 items that were constructed to parallel directly the content of the SPIB-T. Testing and ratings were both accomplished within a three-month period.

33 A variety of correlational analyses were performed within each of the three validity studies. Some of the most interesting results were provided by canonical correlations, using the nine SPIB tests on one side of the relationship and the major sub-scales of the criterion instrument on the other side.[34] These analyses produced canonical correlations of .58 in the SPIB predictive validity study,[35] .61 in the SPIB concurrent validity study, and[36] .75 in the SPIB-T concurrent validity study. These results strongly suggest that knowledge and applied performance in the SPIB domains are related to one another.

Discussion of the validity studies.[37] When one considers the findings of the GATB, TOWER, and SPIB studies together, the overall appraisal is far from encouraging. In a tightly controlled study, the relationship between general aptitudes and vocational

Implications

performance did not appear to be very strong. Since the criterion variable involved performance ratings of unknown reliability, it is possible that the low relationship was caused partially by measurement error. On the other hand, it seems quite likely that the general magnitude of the reported relationship is accurate, implying the vocational success requires a great deal more than skill in a number of very general vocational aptitudes. But didn't we really know that all along?

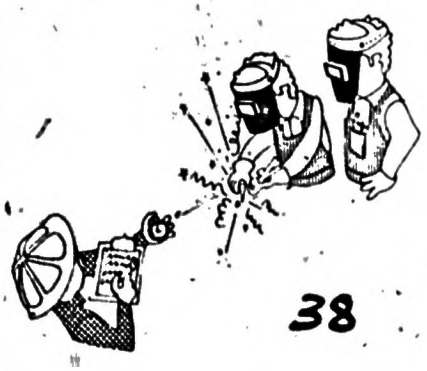
The TOWER project, when it began, must have been seen as providing an opportunity for a major breakthrough. After all, people had been arguing for some time that work samples should provide a much better indication of work performance than scores derived from paper and pencil tests that measure fairly abstract abilities. But the study did not confirm this highly believed hypothesis, and the author was quick to point out the many methodological weaknesses that most certainly contributed to the findings. And so we are left with a question still to be answered by future research. Recent opinion suggests that this question may actually be a dilemma, for it seems to hinge in part on the following issue. Work sample evaluation in the

past has deliberately remained flexible in order to avoid some of the pitfalls of standardized assessment that tend to lower examinee motivation and raise examinee anxiety. The price of this flexibility is increased measurement error, perhaps even to the point of making work sample evaluation unsuitable as a formal testing mechanism.

The SPIB studies are fairly encouraging, showing moderately strong relationships between measures of knowledge and applied performance in several social and prevocational domains. Several methodological factors seem to have facilitated these findings. First, the reliability of both the predictor and criterion variables were ascertained to be satisfactorily high. More importantly, the items for both the predictor and criterion measures, although different in format, were drawn from the same general content domains. This permitted an examination of the relationship between knowledge and applied performance in the same domain, a much more reasonable expectation than knowledge in one domain predicting applied performance in an entirely different domain.

Process or Product of Learning

38 Most psychological measurement in the past has focused only on the products of past learning, assuming



that the opportunities to learn have been adequate. Test items are concerned with what an examinee can do, rather than with what is required to guarantee success. Several researchers in the field of mental retardation, however, have been arguing strongly in recent years that we really ought to be at least equally concerned with measuring the process of learning (Haywood, et al., 1975; Gold, 1973; Budoff & Hamilton, 1976; Kulman, et al., 1975). If our primary question, as Budoff and Hamilton (1976) suggest, is to ascertain a person's ability to learn and profit from optimal instructional experience, then our test formats should include opportunities for examinees to learn and practice what they are expected to perform.

This sentiment has also been strongly expressed by the Vocational Evaluation and Work Adjustment Association (Kulman, et al., 1975). In discussing the strengths and weaknesses of work sample evaluation, they state that testing should always be preceded by some degree of training. In their words, "if a client does not perform adequately following standardized industrial instructions, it is necessary to determine what type(s) of instruction will facilitate his understanding of the task. . . . The evaluation of the clients' ability to learn, their retention, and most efficient means of acquiring information are integral parts of the total assessment process" (p. 59).

Learning Potential

39

39/ Although the research on work sample evaluation does not yet illustrate this principle, Budoff and Haywood in this country and Feuerstein in Israel have conducted several series of studies that have come to be known as "learning potential" research. Their findings strongly suggest that tests which include measurement of learning process will be more predictive of subsequent performance than tests which measure only the current products of past learning. Although these studies have focused primarily upon the measurement of general cognitive abilities, their methods and findings have implications for the more applied domains of vocational and prevocational assessment.

Trainee Performance Sample

40

40/ A recent example of such an application is found in a test called the Trainee Performance Sample (Bellamy & Snyder, 1976). The TPS was designed to predict the practicality of training severely and profoundly retarded adults in a fairly narrow range of light industrial tasks. Notice that the prediction is concerned with practicality rather than feasibility, since the criterion behaviors are known to be feasible. The purpose of testing is to predict the level of resources that will be required to achieve the training objectives.

The TPS consists of 30 items that each represent a learning operation in a vocational context. Testing involves training, including the opportunity to profit from correction. Preliminary reliability and validity data with this instrument are very encouraging.

41/ A more radical approach involving measurement of learning process calls for the full integration of training and evaluation. When this occurs, each step of the training process is monitored, and trainee performance provides immediate feedback concerning the effectiveness of training procedures. If the results are not satisfactory, procedures can be revised until the trainee acquires the desired behaviors (Gold, 1973). As the description of this approach implies, its utility is generally restricted to measurement issues and problems that arise within the context of training.

Training & Evaluation

41

42/ There is a third, perhaps even more radical, context in which the argument for measuring learning process has been promulgated. One example of this position comes from the literature on career education for the handicapped. Dunn (1973) argues that most of our current career education programs provide students with information, rather than teaching them how to find and acquire the information for themselves. Instead

Decision Making

42

of the prevailing approach, Dunn suggests that career education programs should focus on decision-making, which has two components: acquiring information, and developing and implementing a strategy for processing the information. The startling implication of this philosophy is that measurement would focus almost exclusively on the process of learning, drawing upon particular products only in-so-far as they illustrated key components of the process. The work of Goldstein and his associates in development of the Social Learning Curriculum provides an excellent example of this principle.

Norm-Referenced or Criterion-Referenced?

43 One of the most interesting measurement issues to emerge in recent years has been the argumentation in support of criterion-referenced tests as a supplement to, if not a replacement of, norm-referenced tests. Not infrequently, there has been more heat than light shed during presentation of the arguments.[44] The source of the confusion lies in the ambiguity of the word "criterion," which has two completely different meanings. The first meaning, that we have already discussed, refers to the validity of a test in terms

43

CRITERION

Meaning 1 = Ultimate Behavior

Meaning 2 = Standard of Performance

44

of the extent to which the test measures directly or indirectly the criterion behavior of ultimate concern. This meaning of criterion is descriptive. The second meaning of the word "criterion" refers to a critical level or standard of performance. This meaning of criterion is evaluative.

The two major motives that lie in back of the criterion-referenced movement are derived from the two different meanings of criterion (Donlon, 1975). The first motive involves the desire to test people on instructionally relevant dimensions; i.e., the criterion behaviors themselves. The second motive involves the desire to evaluate test scores on the basis of performance magnitude rather than relative position within a group.

45/ Let us consider first the instructional motive. Several quotations are illustrative. Hively (1975) suggests "the most important characteristic of domain-referenced testing is that they [sic] provide students with clear opportunities to try over and over again to achieve well-defined areas of skill. Norm-referenced testing systems do not provide such opportunities". (p. 5). Reynolds (1975) states that "in today's context the measurement technologies ought to become

Meaning 1

Instructional Motive

45

integral parts of instruction designed to make a difference in the lives of children and not just a prediction about their lives" (p. 15).

Both of these quotations allude to the fact that most norm-referenced tests have been designed for the purposes of selection or classification, rather than for the design, modification, or evaluation of instruction. Although this is clearly an historical trend, it is not an inherent necessity. Any test may be constructed in a way that samples criterion behavior in the descriptive sense of this word. Furthermore, selection and classification have their proper place along with the instructional purposes of measurement.

46/ The evaluative meaning of criterion offers some deeper issues for consideration. Any test score has both descriptive and evaluative interpretation. The descriptive aspect of a test is simply its raw score; i.e., the number of items answered correctly. The evaluative aspect of a test is the interpretation placed upon a score indicating level of success. Both criterion-referenced and norm-referenced tests produce raw scores, which might even be derived from the same test items. In one case, evaluation is tied to the performance of one or more reference groups.

Meaning 2

Evaluative Motive

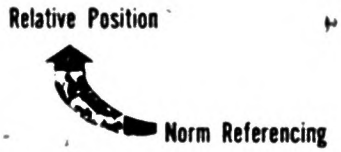
46

In the other case, evaluation is based upon an arbitrary judgment (Donlon, 1975).

Neither approach presents the entire picture, or, stated more positively, each approach presents part of the picture.[47] The main virtue of performance magnitude, as an evaluative dimension, is its direct interpretability as achievement. An examinee who responds correctly to 95% of the items on a test may have performed well. But how adequate is 90%? Or 85%? Inevitably, the judgment is somewhat arbitrary if test performance and implicit values are the only frames of reference.

47

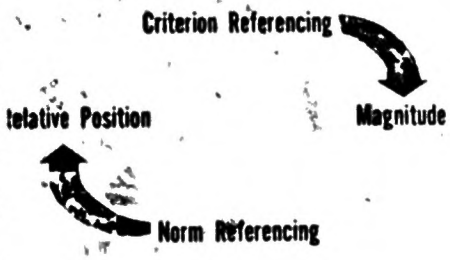
[48] An added perspective may be obtained if an examinee's test scores are compared to the performance of an appropriate reference group. Norm-referencing, however, usually produces an index of relative performance, such as percentiles, which if considered alone, provides no information at all about magnitude. Furthermore, if one reports test scores only in relative terms, this may actually mask the measurement of growth. An individual may grow in magnitude over the years, but never improve his relative position with reference to a norm group. Growth in this context goes unnoticed, which leads one to agree with Donlon (1975) when he states that "from an individual



48

point of view, a great deal is wrong with a number that tells you not how much you did, but how many you outdid" (p. 33).

49/ Given these pitfalls with both approaches to evaluating test scores, it seems worthwhile to construe criterion referencing as an attempt to stipulate arbitrarily the ideal level of performance, and norm referencing as providing an empirical checkpoint for the reasonableness of our expectations. In order for norm referencing to serve such a role adequately, test scores within the reference group will have to be reported in terms of magnitude as well as relative position.



49

50/ There is one other issue that should also be mentioned briefly. Some of the proponents of criterion-referenced testing have suggested that this approach encourages teaching to the test item, a practice almost always discouraged in norm-referenced testing. Upon closer examination, however, we discover that the issue has nothing to do with criterion referencing or norm referencing, but rather is concerned with the breadth of the domain being tested, and the generalizability of an examinee's performance (Donlon, 1975; Rosner, 1975; Hofmeister, 1975; Joselyn, 1975). Rosner states this succinctly when he points out that a test

Teaching to the test?

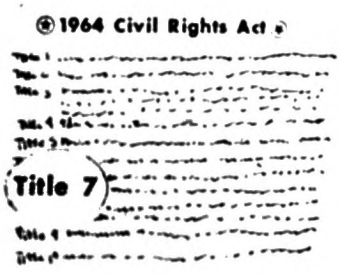
50

"must assess a generalizable behavior -- a skill that transfers within a specific domain. . . Teaching to the test is a worthwhile enterprise only if what is learned will be apparent in other situations which are similar to the test in certain ways but involve different stimuli and/or contexts" (p. 45). This concern with generalizability is relevant for both criterion-referenced and norm-referenced tests.

Enter the Politician

51

51 Most of the issues presented thus far have been concerned with the principles and practices of measurement. During the past decade, however, a political dimension has emerged that is likely to rattle more cages than all of our professional issues combined. It seems fitting, therefore, to close this presentation with a discussion of politics.



52

52 It all began with passage of the 1964 Civil Rights Act. Title VII of this Act stipulates that employers, labor unions, and employment agencies may not discriminate on the basis of race, color, religion, sex, or national origin. Of particular relevance to the field of measurement is John Tower's amendment to Title VII, which reads as follows (Koenig, 1974):

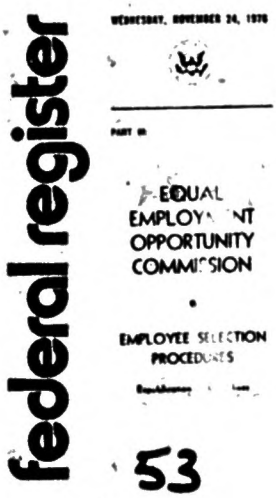
> Nor shall it be an unlawful employment practice for an employer to give and to act upon the results of any professionally developed ability test, provided that such test, its administration or action upon the results, is not designed, intended or used to discriminate because of race, color, religion, sex, or national origin.

In order to implement Title VII, the Equal Employment Opportunities Commission (EEOC) was created.

As Koenig points out (1974), it is not entirely clear whether the purpose of the Tower amendment was to prevent discrimination or to make it legitimate under the disguise of scientific objectivity. In any case, the sale of psychological tests to industry boomed during the late 60's, to the point where some referred to Title VII as the industrial psychologists' Guaranteed Income Act.

53/ In 1970, a somewhat amazing event occurred in the publication of Title VII guidelines by the EEOC (Federal Register, 1970). It was not the mere publication of these guidelines that was amazing, but rather their strength. Consider the following examples:

- Sec. 1607.3. The use of any test which adversely affects hiring, promotion, transfer, or any other employment . . . . opportunity [for minorities] constitutes discrimination unless (a) the test has been validated and evidences a high degree of utility . . . .



and (b) the person giving or acting upon the results of the particular test can demonstrate that alternative suitable hiring, transfer, or promotion procedures are unavailable for his use.

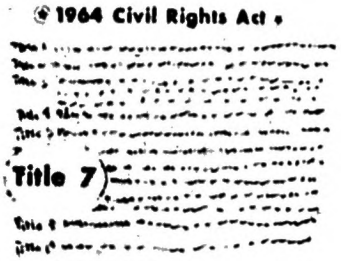
- Sec. 1607.4a. Where technically feasible, a test should be validated for each minority group with which it is used.
- Sec. 1607.4c. Evidence of a test's validity should consist of empirical data demonstrating that the test is predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated.
- Sec. 1607.7. Any person citing evidence from other validity studies as evidence of test validity for his own jobs must substantiate in detail job comparability and must demonstrate the absence of contextual and sample differences [between his applicants and the sample which provided the validity information.]

Who would have believed in 1964 that such stringent guidelines were to be created? Few tests could even come close to complying.

For a while, the EEOC guidelines contained more bark than bite.[54] But on March 8, 1971, the Supreme Court radically changed this picture in its ruling on Griggs vs. the Duke Power Company.

This case involved a class action suit against the Duke Power Company with respect to employment practices at its Dan River station in North Carolina (Koenig, 1974). The company divided its labor force into five divisions: labor, coal handling, operations, maintenance, and laboratory and test. Labor involved basically janitorial work, and was the division to which all but one of the company's 14 black employees were assigned in 1966. Incidentally, the maximum wage paid to a black employee was $1.65 per hour, whereas the entry wage for white employees was $1.88 per hour.

After July 2, 1965, the company decided that black people could be assigned to the coal handling division, provided they held a high school diploma and earned passing scores on the Wonderlic Personnel Form, a test of cognitive ability, and the Bennet Mechanical Comprehension Test. Griggs argued and Supreme Court concurred that these conditions had little to do with ability to shovel coal.



55

55/ The impact of the Griggs decision was to reinforce the validity stipulations of the EEOC guidelines. At this point in time, there is no direct extension of Title VII to employment opportunities for the handicapped. Indeed, there are Sections 503 and 504 of the 1973 Vocational Rehabilitation Act which prohibit federal contractors from discriminating against the handicapped and which generally prohibit discrimination against the handicapped by any recipient of federal funds.[56] Thus far, however, the issues surrounding measurement of handicapped people as a potential source of discrimination have not been translated into regulations. In my opinion, we would all be well advised to deal carefully with these issues as a profession before the impetus, if not the control, is taken away from us by the politicians.

56

Aptitude	Test
G - Intelligence	Three dimensional space Arithmetic reasoning Vocabulary
V - Verbal Intelligence	Vocabulary
N - Numerical Aptitude	Computation Arithmetic reasoning
S - Spatial Aptitude	Three dimensional space
P - Form Perception	Tool matching Form matching
Q - Clerical Perception	Name comparisons
K - Motor Coordination	Mark making
F - Finger Dexterity	Assemble Disassemble
M - Manual Dexterity	Place Turn

Figure 1

Components of the General Aptitude Test Battery

Table 1

Relationship Between Test Norms

(G--105, V--105, N--105) and Dichotomized Criterion--

Case Worker 195.108: N = 106

	Non-Qualifying Test Scores	Qualifying Test Scores	Total
Good Workers	20	54	74
Poor Workers	18	14	32
TOTAL	38	68	106

$\phi = .28$

$P/2 < .005$

References

Bellamy, G. T., & Snyder, S. The trainee performance sample: Toward the prediction of habilitation costs for severely handicapped adults. In T. Bellamy (Ed.), Habilitation of severely and profoundly retarded adults. Eugene, Oregon: Specialized Training Program, University of Oregon, 1976, 79-90.

Budoff, M., & Hamilton, J. Optimizing test performance of moderately and severely mentally retarded adolescents and adults. American Journal of Mental Deficiency, 1976, 81(1), 49-57.

Donlon, T. Referencing test scores: Introductory concepts. In W. Hively & M. Reynolds (Eds.), Domain referenced testing in special education. Reston, Va.: Council on Exceptional Children, 1975, 29-42.

Federal Register. Guidelines on employee selection procedures. 35FR12333, August 1, 1970.

Gold, M. Research on the vocational habilitation of the retarded: The present, the future. In N. Ellis (Ed.), International review of research in mental retardation: Volume 6. New York: Academic Press, 1973, 97-148.

Halpern, A., Raffeld, P., Irvin, L., & Link, R. Measuring social and prevocational awareness in mildly retarded adolescents. American Journal of Mental Deficiency, 1975, 80, 81-89. (a)

Halpern, A., Raffeld, P., Irvin, L., & Link, R. Social and prevocational information battery. Monterey, Ca.: CTB/McGraw-Hill, 1975. (b)

Haywood, H. C., et al. Behavioral assessment in mental retardation. In P. McReynolds (Ed.), Advances in psychological assessment: Volume 3. Washington: Jossey-Bass, 1975, 96-136.

Hively, W. Introduction. In W. Hively & M. Reynolds (Eds.), Domain-referenced testing in special education. Reston, Va.: Council on Exceptional Children, 1975, 1-14.

Hofmeister, A. Integrating criterion-referenced testing and instruction. In W. Hively & M. Reynolds (Eds.), Domain-referenced testing in special education. Reston, Va.: Council on Exceptional Children, 1975, 77-88.

Irvin, L., & Halpern, A. Reliability and validity of the Social and Prevocational Information Battery for mildly retarded individuals. American Journal of Mental Deficiency, In press.

Irvin, L., Halpern, A., & Reynolds, W. Social and Prevocational Information Battery - Form T. Eugene, Oregon: Rehabilitation Research and Training Center, University of Oregon, 1977.

Irvin, L., Halpern, A., & Reynolds, W. Measuring social and prevocational awareness in moderately retarded individuals. American Journal of Mental Deficiency, In press.

Joselyn, G. Ethical considerations in the use of standardized tests. In W. Hively & M. Reynolds (Eds.), Domain referenced testing in special education. Reston, Va.: Council on Exceptional Children, 1975, 121-140.

Koenig, P. They just changed the rules on how to get ahead. Psychology Today, June, 1974, 87-95, 100-103.

Kulman, et al. The tools of vocational evaluation. Vocational Evaluation and Work Adjustment Bulletin, 1975, 8, 49-64.

Lavisky, S. Invited address. In J. Sanders & T. Sachse (Eds.), Problems and potentials of applied performance testing. Portland, Oregon: Northwest Regional Educational Laboratory, 1975, 33-54.

McClelland, D. Testing for competence rather than for 'intelligence.' American Psychologist, 1973 28, 1-14.

Manual for the USES General Aptitude Test Battery. Section III: Development. United States Department of Labor, Manpower Administration, 1970.

Reynolds, M. Trends in special education: Implications for measurement. In W. Hively & M. Reynolds (Eds.), Domain-referenced testing in special education. Reston, Va.: Council on Exceptional Children, 1975, 15-28.

Rosenberg, B. The job sample in vocational evaluation. Final Report of Project RD-561. New York: Institute for the Crippled and Disabled, 1967.

Rosner, J. Testing for teaching in an adaptive educational environment. In W. Hively & M. Reynolds (Eds.), Domain-referenced testing in special education. Reston, Va.: Council on Exceptional Children, 1975, 43-76.

Testing orientation and work evaluation in rehabilitation. ICD Rehabilitation and Research Center. New York: 1974.